CS-IRISH WHISKEY IMPORT FORCAST Rubric

DS 4002 – Spring 2023 – Lyle Johnson
Due: One week after assigned
Submission format: Upload link to github repo

Individual Assignment

General Description: Submit to canvas a link to your case study repository.

Preparatory Assignments – Everything in the course.

Why am I doing this? As a data scientist, it's important for you to have experience working through various research problems across various fields you may or may not be familiar with. While data science tools are universal, it can be uncomfortable to be thrown into a problem whose context you have little prior experience with. In these situations, it's important to stay calm and not rush into producing results. Attack the problem diligently through research; turn to your piers who may be more experienced with the topic at hand; ask questions and don't be afraid to try out new things, even if they fail at first. In completing this case study and pushing yourself outside of your comfort zone, you will become a more versatile, confident data scientist.

- Course Learning Objective: establish data sets relevant to your hypothesis/model
- Course Learning Objective: create a functioning data science pipeline
- Course Learning Objective: prepare findings for presentation to your peers

What am I going to do? First, read the one-page hook document to familiarize yourself with the project context, problem, and overall objective. Regardless of whether you have experience forecasting with time series data, brainstorm how you would conceptually approach the problem while reading through the document. Then, turn to the datasets in the github repo for this case study and familiarize yourself with the current data format. Afterwards, follow the guidelines below to produce the following:

- Github repository containing:
    - Finalized dataset in excel format, properly cleaned and ready for analysis
    - One code file containing code chunks for stationary testing, autocorrelation testing, and forecasting by region for the next three years
    - One-page PDF document assessing stationarity and autocorrelation test results
    - One-page PDF document assessing forecasting model results and making a recommendation as to whether your model is an appropriate model for your company to base critical business decisions on

- README document orienting users to your repo and including any additional resources used

All of this will be submitted electronically via a link to a github repository

Tips for success:

- Take your time. Read all supporting materials through before starting to clean data or run analysis.
- Talk to the professor and the TA. You may not have experience working with time series data before – don't be afraid to ask questions if you're stuck.
- Confidently make a decision. Once forecasting is complete, decide whether you think your forecast should be used as a legitimate indicator of future whiskey sales. Make a recommendation, explain your reasoning thoroughly, and stick to your guns.

How will I know I have Succeeded? You will meet expectations for the Case Study when you follow the criteria in the rubric below.

Spec Category	Spec Details

Formatting	• Repository – A github repo containing all materials
	  ○ Submit a link to the repo
	  ○ Everything is contained in the repo or linked to it if appropriate
	  ○ **Contents**:
	  ▪ Final dataset in excel format
	  ▪ Source code file
	  ▪ PDF document 1
	  ▪ PDF document 2
	  ▪ README.md document

Dataset	• **Goal**: This excel file should contain your final dataset as used for forecasting • Using the "Imports" and "World_Bank" files in the case study repo, create one final dataset which merges import data for each country with income and region data • Final dataset rows should be unique whiskey import data for each country for each year • Final dataset should include columns for *Quality, Country, Income, Region, Year,* and *Cases*. • Drop any rows whose *Name* is preceded by "DF"
Source code file	• **Goal**: This file should contain code for conducting stationary testing, autocorrelation testing, and forecasting by region for the next three years • Stationarity, autocorrelation, and forecasting analysis must be conducted for each region in the dataset; they may be conducted using any appropriate tests/models, but supplemental documents in the case study repo provide guidance as to which methods may be appropriate

	• Code may be written in preferred coding language and file uploaded in any appropriate format
PDF document 1	• **Goal**: One-page PDF document assessing stationarity and autocorrelation test results • Explain the specific tests you chose to use and justify your decision to use theses tests • Justify your decision regarding the lag value used for your autocorrelation tests • Include explicit conclusions as to whether time series data for each region is stationary or not, and highly autocorrelated or not – explain what these conclusions mean for forecasting analysis • Explain how results from this section influence your decision regarding the proper forecasting model to use

PDF document 2	• **Goal**: One-page PDF document assessing your forecasting results and making a recommendation to your boss as to whether you believe your model is an appropriate model for your company to base critical business decisions on
	• Forecast Irish whiskey imports for each region for the next 3 years
	• Plot results on one time series plot and include this plot in your PDF
	• Explain the specific forecasting model you chose to use and justify your decision to use this model
	• Interpret your forecast results in a business context, i.e., if your forecast is correct, what should your company do?
	• Make a recommendation to your boss as to whether she should actually base business decisions on your forecast
	  ○ It may be helpful (though not necessary) to include a discussion of model complexity, data quality, and other considerations in this section
	  ○ It does not matter if you support or reject your forecast results, only that you defend your decision with evidence

README.md	• **Goal**: This file provides an overview of the contents of your repo and orients visitors
	• Use markdown headers to divide content
	• Format:
	  ○ Data section
	  ▪ Data dictionary
	  ○ Source code section
	  ▪ Explain usage of code – i.e., if/what edits users need to make to run the code
	  ○ References section
	  ▪ Include properly formatted (IEEE style) references to any additional sources used outside of resources already provided in this case study

Acknowledgements: Special thanks to Jess Taggart from UVA CTE for coaching on making this rubric. This structure is pulled direction from Streifer & Palmer (2020).